Työmaana tutkimusdatan avoimuus – Yhteiskuntatieteellisen tietoarkiston näkökulma

Johtaja Sami Borg, Yhteiskuntatieteellinen tietoarkisto

Tutkijapalvelut näkyviksi – tutkimusaineistot ja kirjaston rooli
STKS:n seminaari, Helsinki, Tieteiden talo, 4.11.2013

YHTEISKUNTATIETEELLINEN TIETOARKISTO
FINLANDS SAMHÄLLSVETENSKAPLIGA DATAARKIV
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Esityksen rakenne

- Taustaa, tutkimusdata, avoin data
- Työmaat: Yhteiskuntatieteellinen tietoarkisto ja datan avaaminen
- Kirjastot ja tutkimusdataan liittyvät palvelut

Taustaa 2013

- Tutkimusdatan avaaminen yleistyy, monia syitä (tietotekniset edistysaskeleet, open access –aate, toimijoiden lukumäärän ja osaamisen kasvu, tarve tutkimusjärjestelmien tuottavuuden lisäämiseen, kilpailu, uutuusarvo jne.)
- Datan avoimuutta tukevat kansainväliset suositukset ja päätökset (OECD, EU)
- Kansallisten tutkimusrahoittajien datapolitiikat (Yhdysvallat: NSF ym.; UK/Research Councils, EU, SA)
- ESFRI-prosessi ja tutkimusinfrastruktuurien kehittäminen
- Suomessakin paljon kansallisia infrastruktuureja ja hankkeita, joilla ESFRI-kytkentä (SSH-alalla esim. FSD/CESSDA, CLARIN, ESS)
- OKM: CSC/Tutkimuksen tietoaineistot TTA

Tutkimusaineistojen avoimuus Suomessa / yhtymäkohtia päätöksiin, suosituksiin ja toimiin

- Euroopan komission digitaalistrategia (2011-)
- Riding the wave. How Europe can gain from the rising tide of scientific data (EU 2010)
- Komission suositus tieteellisen tiedon saatavuudesta ja säilyttämisestä (EU 2012)
- Access to Research Data from Public Funding (OECD 2007)
- Kataisen hallituksen ohjelman tietovarantomaininnat (2011–2015)
- Valtioneuvoston periaatepäätös avoimen datan lisäämisestä (2011)
- EU:n tietosuoja-asetus (2014)
- Tilastolaki, Arkistolaki, Tietohallintolaki
- ESFRI-prosessi ja tutkimusinfrasruktuurien tiekartan päivitys (2013-)
- SSH-ESFRI-hankkeiden eteneminen (esim. CESSDA ja Fin-CLARIN)
- Suomen Akatemian FIRI-asiantuntijaryhmä (2012-)
- Tutkimusinfrastruktuurien kansallisen tiekartan päivitys (2013-)
- Opetus- ja kulttuuriministeriön infrastruktuurihankkeet: a. Tutkimuksen tietoaineistot- ja TTA-hanke - Olennaisen käsikirja päättäjille (2010) - Tieto käyttöön -raportti (2011) b. Kansallinen digitaalinen kirjasto KDK

Tutkimusdata?

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007):

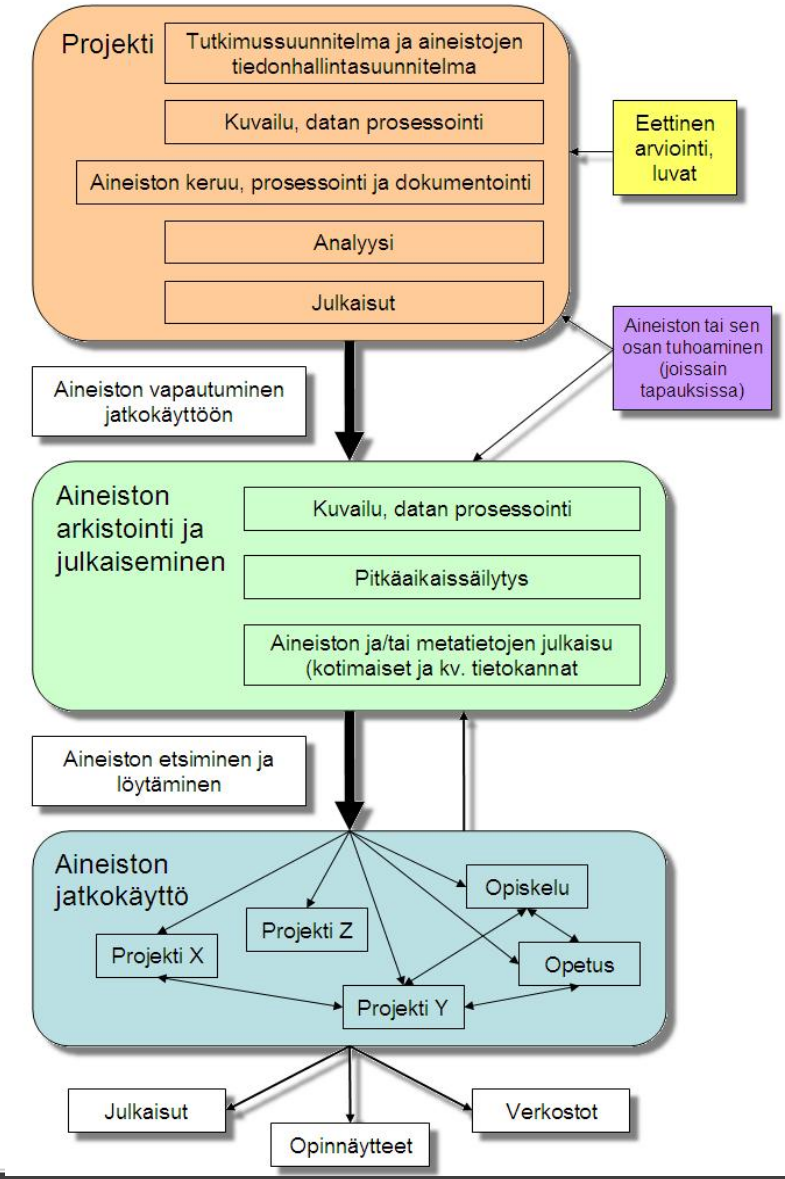
In the context of these Principles and Guidelines, “research data” are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are

commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.

Avoin data? →Avoimen datan hyötykäytön ulottuvuuksia

- Tietosisällön laatu → käytön yleinen mielekkyys, yleinen soveltuvuus uusiin käyttötarkoituksiin
- Yleinen tekninen käytettävyys / käytön tehokkuus: koneluettavuus, yhteensopivuus
- Sisällön käytettävyys: dokumentointi, tarkistaminen, muokkaus, korjaaminen → virheettömyys, eheys, tarkkuus, läpinäkyvyys
- Saatavuuden helppous: helppo saavutettavuus (esim. www), maksuttomuus tai riittävä edullisuus
- Käytön esteettömyys: käyttörajoituksettomuus tai korkeintaan vähäisiä käyttörajoituksia; kuitenkin käyttölisenssit ja käyttömahdollisuus lähdeviittausvelvoittein
- Yhdistettävyys ja linkitettävyys (datan harmonisointi, avainmuuttujat jne.)
- Open data: innovaatiot ja kaupallinen hyödynnettävyys korostuvat

Tietoarkisto ja tutkimusdatan elinkaari



FSD:n palvelujen hyödyntäminen

- Tutkimusaineistoja tietoarkistoon tallentaneita organisaatioita yhteensä 120
- Aineistovarannossa lähes 1400 tutkimusaineistoa (surveyt + kvalitatiiviset tekstiaineistot, joitakin AV-aineistoja)
- Vuonna 2012 tutkimusaineistoja tilattiin noin 700 (käyttölupahakemusten perusteella), yhteydenotot FSD:stä tai FSD:hen yht n. 3100 → keruuseen ja käyttöön liittyvä tietopalvelu >
- Kaikkiaan palvelujen käyttäjiä vuosittain yli 2000 , ulkomaisen käytön osuus toistaiseksi noin 15-20 %
- Alusta pitäen käytössä laajassa kansainvälisessä yhteistyössä kehitetty metadataformaatti DDI (Data Documentation Initiative)
- FSD:n aineistojen sähköinen tilaus- ja toimitusjärjestelmä valmistuu vuoden 2013 loppuun mennessä

Kokemuksia infrastruktuurin perustamisesta ja toiminnasta

- Tutkimusdatan avoimuutta kannatetaan – periaatteessa
- Toimintakulttuurit muuttuvat hitaasti
- Tutkijat haluavat ja heidän pitää saada keskittyä tutkimukseen
- Palvelujen käyttö edellyttää aktiivista tiedottamista ja vuorovaikutusta
- Tutkimusaineistojen hallinnan ja elinkaaresta huolehtimisen vaatimukset kasvavat
- Tutkijat eivät osaa eivätkä aina myöskään halua käyttää tietopalveluammattilaisten palveluja

- Eri aloille tarvitaan sitoutuvia datapalvelujen ammattilaisia

Sami Borg ja Arja Kuula

Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari

Valmisteluraportti OECD:n datasuosituksen toimeenpanomahdollisuuksista Suomessa

Yhteiskuntatieteellisen tietoarkiston julkaisuja 6, 2007

Suosituksen mukaan tutkimusaineistojen avoin saatavuus ja tutkimusaineistojen jakaminen avoimeen käyttöön edistää merkittävästi

- tieteen avoimuutta
- vaihtoehtoisia tutkimusasetelmia
- tiedonkeruun ja tutkimusmenetelmien tutkimusta
- tutkijakoulutusta
- uusien, aineistojen kerääjiltä tutkimatta jääneiden aiheiden tutkimusta
- erilaisten aineistojen ja tietojen liittämistä toisiinsa.

Datasuositus tukee periaatteellisella tasolla

- tutkimusaineistojen avoimuutta tukevia toimintakulttuureja
- tutkimusaineistojen avoimuutta edistäviä hyviä käytäntöjä
- tutkimusaineistojen avoimuuden hyötyjen ja haittojen avointa tiedottamista
- jäsenmaiden tiedepoliittisia toimenpiteitä tutkimusaineistojen avoimuuden edistämiseksi
- suosituksia, joilla tuetaan tutkimusaineistojen avoimuutta kansainvälisessä toimintaympäristössä.

Mitä tutkimusdatoja suositus koskee ja miten?

- Ensisijaisesti julkisrahoitteisia, *sähköiseen* muotoon tallennettuja, tutkimustulosten pohjaksi ja validoimiseksi koottuja tutkimusaineistoja.
- Sähköinen tutkimusdata voi koostua numero-, teksti-, kuva- tai äänitallenteista.
- Tutkimusdata on usein tutkimuksen ns. pohja- tai perusaineisto.
- Suosituksen piiriin kuuluu suuri osa yliopistoissa, valtion ja muun julkisen sektorin tutkimuslaitoksissa sekä muiden julkisen sektorin tiedontuottajien piirissä kerättävistä tai näiden tahojen hallinnoimista tutkimusaineistoista.
- Suosituksen periaatteet ovat useissa tapauksissa sovellettavissa muihinkin kuin julkisrahoitteisesti koottuihin sähköisiin tutkimusaineistoihin.

Mitä avoin saatavuus (Open Access) tarkoittaa sähköisten tutkimusaineistojen osalta?

- Keskeisten julkisrahoitteisten tutkimusdatojen tulisi lähtökohtaisesti olla avoimessa käytössä.
- Mahdollisimman pienin kokonaiskustannuksin turvataan mahdollisimman suurelle käyttäjäkunnalle esteetön ja tasavertainen tutkimusdatan käyttömahdollisuus.
- Tyypillisesti kyse on tutkimusdatan kerääjän ensikäytön tai muun aktiivikäytön jälkeisestä jatkokäytöstä.
- Suurissa aineistoinfrastruktuureissa kyse voi olla myös ensikäyttöön verrattavasta käyttömahdollisuudesta, joka perustuu usein laajaan yhteisrahoitukseen ja -käyttöön.
- Tutkimusdatojen Open Access ei aina tarkoita täysin avointa ja maksutonta tutkimusdatan (jatko)käyttöä.
- Avoimen saatavuuden lisäksi suosituksessa kiinnitetään paljon huomiota tutkimusdatan informoidun jatkokäytön turvaamiseen (asianmukaiseen ja riittävään tietoon perustuvaan käyttöön).
- Tutkimusdatan avoimen saatavuuden lisääminen edellyttää usein tutkimusaineistojen dokumentointia ja muokkausta, joilla aineisto saatetaan käyttökuntoon jatkokäyttöä varten.
- Jatkokäyttö voi sisältää käyttörajoituksia ja avoimenakin edellyttää erilaisia käyttölupia, käyttöehtositoumuksia ja käyttäjätunnistusta.
- Aineistojen saatavuutta voidaan edistää tehokkaasti tietoverkkojen avulla.

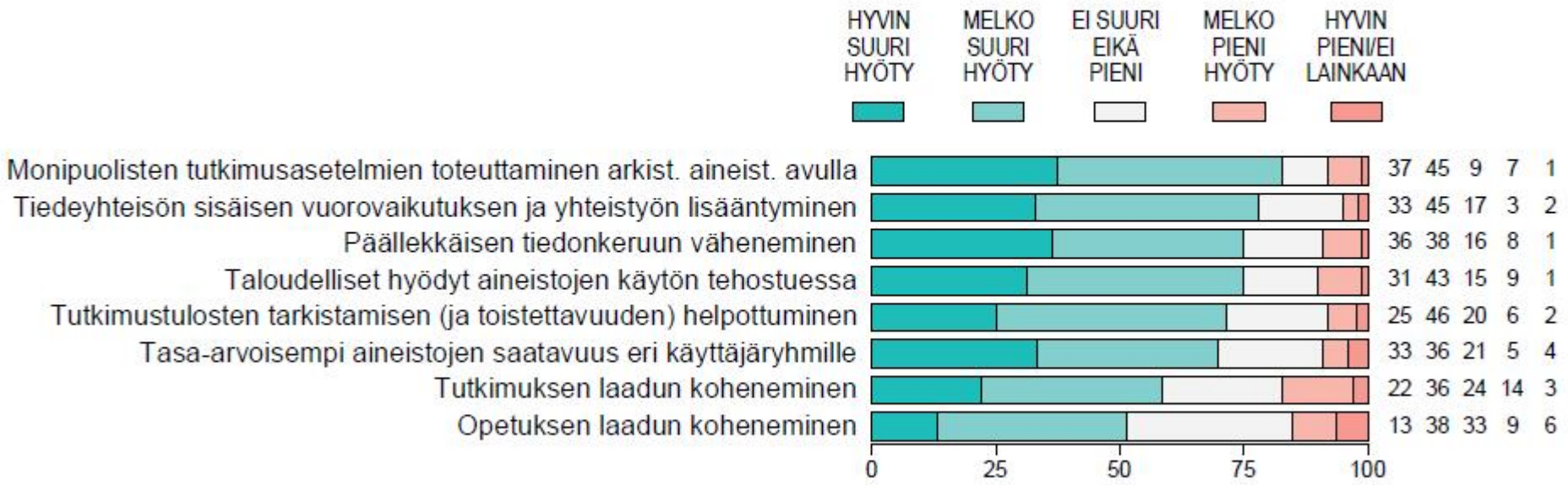
Verkkokysely suomalaisille KY-alan professoreille, marras-joulukuu 2006

- Vastaajina 150 professoria, vastausprosentti 28,4

Aineiston viittaustiedot

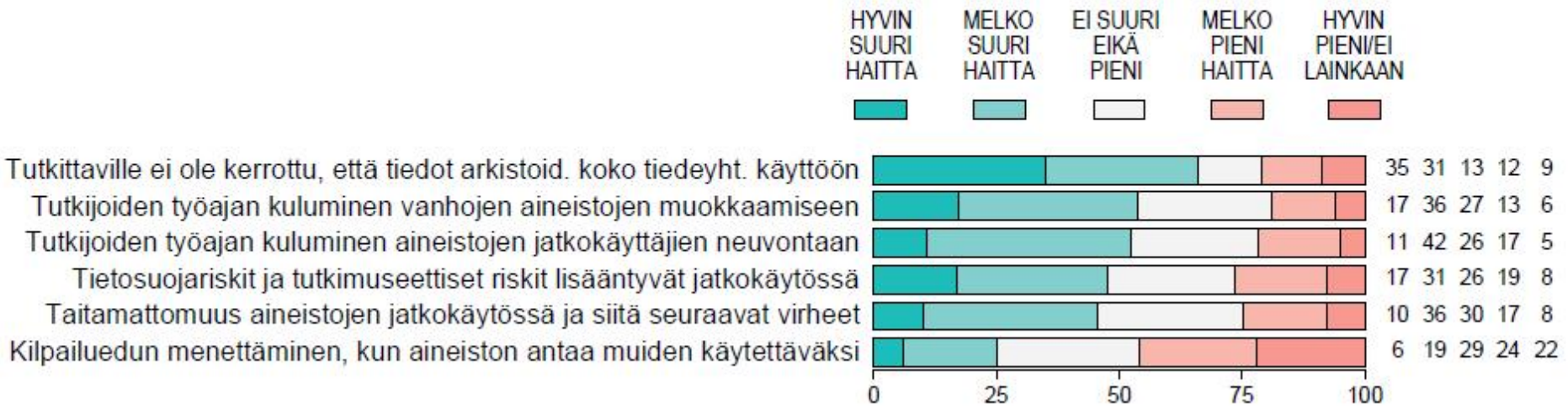
- Borg, Sami & Kuula, Arja: Tutkimusaineistojen säilytys ja avoin saatavuus 2006 [elektroninen aineisto]. FSD2268, versio 1.0 (2007-07-30). Tampere: Yhteiskuntatieteellinen tietoarkisto [jakaja], 2007

Kuvio 4.4. KUINKA SUURINA TAI PIENINÄ PITÄÄ AVOIMUUDEN LISÄÄMISEN HYÖTYJÄ ERI ASIOISSA (%).



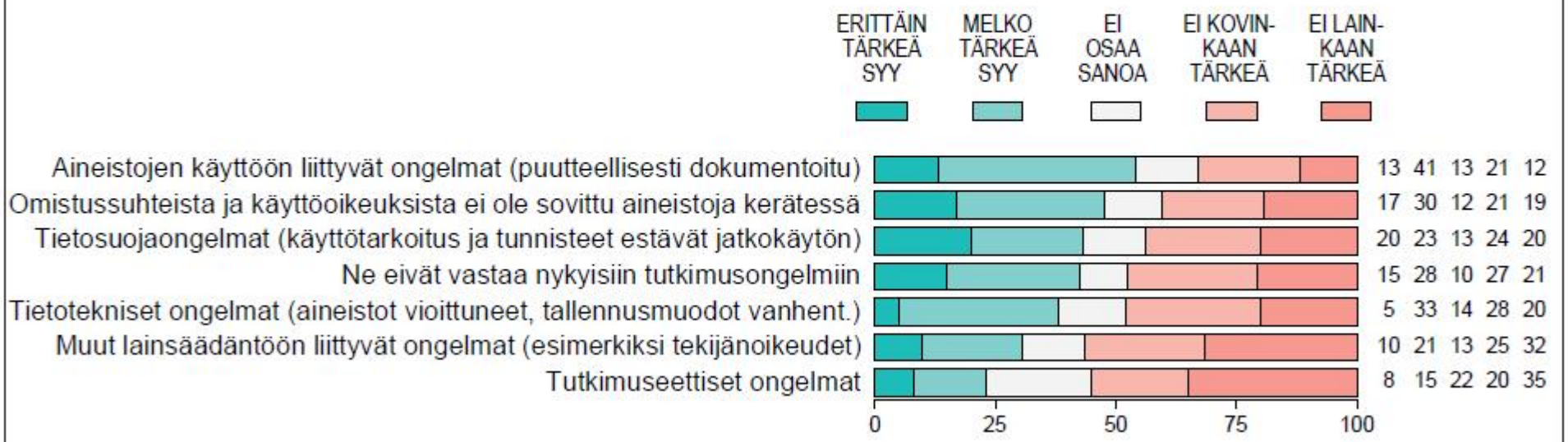
Sähköisen tutkimusdatan avoimuus / Valmisteluraportti OECD:n datasuosituksesta (2007)

Kuvio 3.4. KUINKA SUURINA TAI PIENINÄ PITÄÄ AVOIMUUDEN LISÄÄMISEN MAHDOLLISIA HAITTOJA ERI ASIOISSA (tekstejä osin lyhennetty, %).



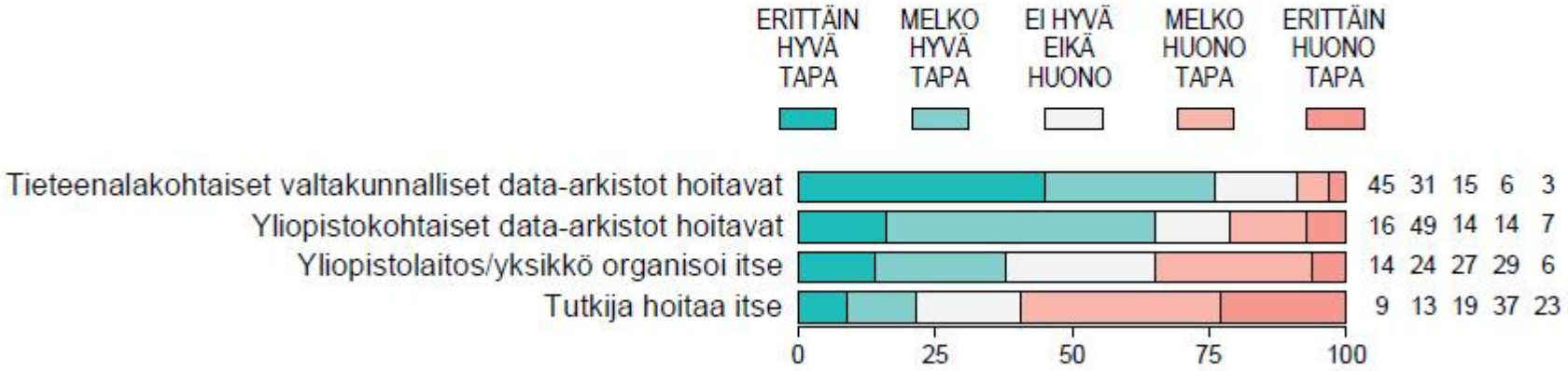
Sähköisen tutkimusdatan avoimuus / Valmisteluraportti OECD:n datasuosituksesta (2007)

Kuvio 3.3. KUINKA TÄRKEITÄ ERI SYYT OVAT SILLE, ETTÄ JO PÄÄTTYNEIDEN TUTKIMUSTEN SÄHKÖISIÄ AINEISTOJA EI JATKOKÄYTETÄ OMALLA TUTKIMUSALALLA (%).



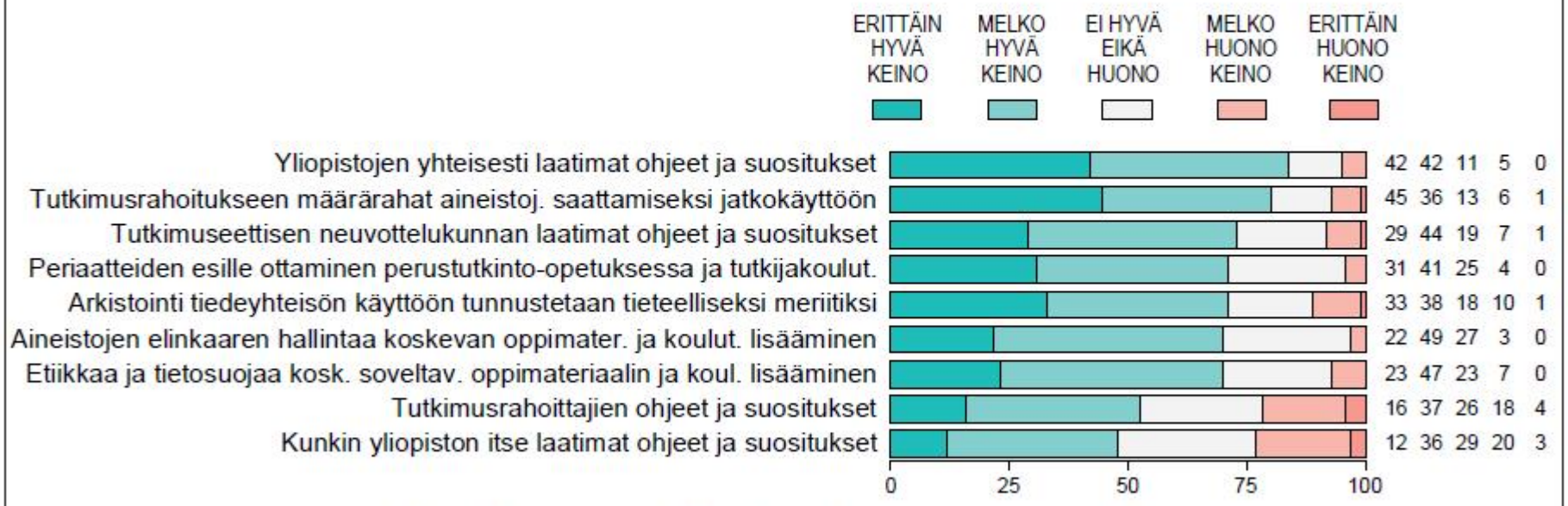
Sähköisen tutkimusdatan avoimuus / Valmisteluraportti OECD:n datasuosituksesta (2007)

Kuvio 5.1. MITEN HYVINÄ TAI HUONOINA PITÄÄ ERI TAPOJA TOTEUTTAA SÄHKÖISTEN TUTKIMUSAINEISTOJEN ARKISTOIMINEN JA SIIHEN LIITTYVÄT TOIMINNOT (kysymys kokonaisuudessaan, ks. teksti; %).



Sähköisen tutkimusdatan avoimuus / Valmisteluraportti OECD:n datasuosituksesta (2007)

Kuvio 6.3. KUINKA HYVIÄ TAI HUONOJA ERI KEINOT OLISIVAT TUTKIMUSAINEISTOJEN OPEN ACCESS -PERIAATTEIDEN TOIMEENPANEMISEKSI (tekstejä osin lyhennetty, %).



Sähköisen tutkimusdatan avoimuus / Valmisteluraportti OECD:n datasuosituksesta (2007)

Tutkimusdatan hallintaa ja avoimuutta edistäneitä toimia Suomessa 2008-

- Tutkimusinfrastruktuurien kansallinen tiekartta, FIRI
- Tutkimuksen tietoaineistot -hankkeet 2009-2010 ja 2011-2013 (OKM / CSC / muut toimijat)
- Kansallinen digitaalinen kirjasto
- SA ryhtynyt edellyttämään kaikilla aloilla aineistonhallintasuunnitelmaa osana tutkimussuunnitelmia (Useat muut rahoittajat ovat seuranneet tai seuraavat perässä)
- TENK / tutkijan ansioluettelomalli → meritoituminen myös tutkimusdatan julkaisemisesta
- Kansainväliset suositukset (OECD, EU) ja yleinen kehitys

Tutkimuksen tietoaineistot -hanke TTA

- Opetus- ja kulttuuriministeriön rahoittama hanke vuosille 2011–2013. Hankeen toteuttajana toimii CSC.
- Hanke panostaa tutkimuksen tietoaineistojen hyödyntämiseen liittyvän tahtotilan vahvistamiseen ja kansallisen tietopolitiikan luomiseen sekä tutkimuksen tietoinfrastruktuurin rakentamiseen.
- Hankkeen aikana mm. rakennetaan federoitu tutkimusaineistojen tallennuspalveluratkaisu, tuetaan metatiedon tuottamista ja yhdenmukaistetaan tutkimusaineistojen tuottamiseen ja ylläpitämiseen liittyviä prosesseja sekä selvitetään Kansallisen digitaalisen kirjaston kanssa yhteinen pitkäaikaistallennusratkaisu.
- Hankkeen taustalla on vuosina 2009–2010 käynnissä ollut Tutkimuksen tietoaineistot -selvityshanke sekä sen pohjalta vuonna 2011 valmistunut selvitys Tieto käyttöön. Tiekartta tutkimuksen sähköisten tietoaineistojen hyödyntämiseksi
- TTA jatkuu vuonna 2014 osana CSC:n toimintaa

TTA: IDA

”IDA on turvallinen ja helppokäyttöinen säilytyspalvelu datalle ja siihen liittyvälle metatiedolle. IDA avattiin käyttöön 20.9.2012. IDA-palvelussa tarjotaan 2017 loppuun saakka noin viiden petatavun (PT) tallennuskapasiteetin käyttöoikeuksia.

IDA on tutkimusjärjestelmän toimijoille suunnattu tutkimusdatan säilytyspalvelu, jonka käyttöperiaatteet asettaa opetus- ja kulttuuriministeriö.”

Lähde: http://www.tdata.fi/ida [22.10.2013]

TTA: KATA

- Kata on tutkimusaineistojen kuvailuja eli metatietoja keräävä palvelu, jota toteutetaan osana opetus- ja kulttuuriministeriön TTA-hanketta.
- Aineistojen löytymisen takaamisen lisäksi Katan tarkoituksena on tuottaa tietoa aineistojen olemassaolosta rahoittajille, mahdollistaa yhtenäisen käyttöehto- ja käyttöoikeuskulttuurin luominen sekä auttaa tunnistamaan ja löytämään tietoaineistoja pitkäaikaissäilytykseen.
- Metatiedoista löytyvät nimet, tekijät, kuvaukset, asiasanat, käyttöehdot ja yhteystiedot. Katan metatiedot ovat vähintään TTA -projektin minimimetatietojen mukaisia. Yhteiset metatiedot on suunniteltu niin, että ne pätevät mihin tahansa aineistoon eli niissä ei ole liikaa tieteenalakohtaista tyyppiriippuvuutta ja kaikille on samat kuvaus- ja asiasanat.

Lähde: http://www.csc.fi/csc/ajankohtaista/uutiset/kata_pilotit [22.10.2013]

TTA: PAS

” Tutkimusaineistojen pitkäaikaissäilytysratkaisua (TTA-PAS) valmistellaan parhaillaan, ja palvelu tullaan ottamaan käyttöön todennäköisesti vuonna 2015.

Pitkäaikaissäilytyspalvelun kehittäminen liittyy Tutkimuksen tietoaineistot (TTA) - ja Kansallinen digitaalinen kirjasto (KDK) - hankkeissa valmisteltavaan pitkäaikaissäilytysjärjestelmään (TTA-PAS ja KDK-PAS), jonka tavoitteena on vähentää digitaalisten kulttuuriperintöaineistojen ja tutkimuksen tietoaineistojen hallinnan, jakelun ja säilyttämisen päällekkäisiä toimia ja pidemmällä aikavälillä kustannusten nousua.

Yhteinen digitaalisten aineistojen pitkäaikaissäilyttämisen ratkaisu tuottaa korkealaatuisia palveluita kustannustehokkaasti ja mahdollistaa digitaalisten aineistojen monipuolisen hyödyntämisen. ”

Lähde: http://www.tdata.fi/pas [22.10.2013]

TTA: REMS

- ” TTA:n käyttövaltuuspalvelu (nimi julkistetaan myöhemmin) perustuu REMSiin (Resource Entitlement Management System).
- Palvelu on sähköinen väline tutkimusaineistojen ja –resurssien käyttövaltuuksien hallintaan. Aineiston käyttölupaa hakevat tutkijat kirjautuvat välineeseen kotiorganisaationsa käyttäjätunnuksella ja salasanalla, täyttävät aineiston sähköisen käyttölupahakemuksen ja sitoutuvat aineiston käyttöehtoihin.
- REMS-väline kierrättää käyttölupahakemuksen hyväksyttäväksi aineiston omistajalle tai hänen nimeämälleen edustajalle. Lisäksi REMS tuottaa tarvittavat raportit hakemuksista ja myönnetyistä käyttöoikeuksista.”
- Lähde: http://www.tdata.fi/rems [22.10.2013]

Kirjastojen rooli?

- Tieteellisten kirjastojen merkitys tutkimusdataa koskevissa asiantuntijapalveluissa kasvaa
- Useita mahdollisia kehityspolkuja ja rooleja: luultavasti kirjastojen palvelumallit tulevat eriytymään
- Malli 1: Keskeisiä tutkimusdatoja ja datapalveluja koskeva tietopalvelu (Mitä on saatavilla mistä?)
- Malli 2: Myös tutkimusdatan saatavuuteen ja käyttöön liittyvät yksityiskohtaiset asiantuntijapalvelut, ei datan ”houstausta” (Edellisen lisäksi käyttöehtoihin ja käyttöön liittyvä sisällöllinen ja tekninen tuki, mutta ei datan hoiva-, säilytys- ja jakelupalveluja)
- Malli 3: Tutkimusdatan elinkaaripalvelut (keruun tuki, datan dokumentointi ja arkistointi, muut hoivapalvelut, pitkäaikaissäilytys, jakelu, käyttöön liittyvä neuvonta)
- → kirjastojen datapalveluvalikko hotellitermein: ei aamiaista, aamiainen, puolihoito, täysihoito
- Yhteensovittaminen jo rakennettuihin datapalveluihin:FSD, CSC jne.

Yksi toimintamalli→Data Library

- "A **data library** refers to both the content and the services that foster use of collections of numeric, audio-visual, textual or geospatial data sets for secondary use in research"
- "The data library tends to house local data collections and provides access to them through various means (CD-/DVD-ROMs or central server for download)."
- "A data library may also maintain subscriptions to licensed data resources for its users to access. Whether a data library is also considered a data archive may depend on the extent of unique holdings in the collection, whether long-term preservation services are offered, and whether it serves a broader community (as national data archives do)."
- Source: http://en.wikipedia.org/wiki/Data_library [2013-10-21]

Services offered by data libraries and data librarians

- Library service providing support at the institutional level for the use of numerical and other types of datasets in research. Amongst the support activities typically available:
- Reference Assistance — locating numeric or geospatial datasets containing measurable variables on a particular topic or group of topics, in response to a user query.
- User Instruction — providing hands-on training to groups of users in locating data resources on particular topics, how to download data and read it into spreadsheet, statistical, database, or GIS packages, how to interpret codebooks and other documentation.
- Technical Assistance - including easing registration procedures, troubleshooting problems with the dataset, such as errors in the documentation, reformatting data into something a user can work with, and helping with statistical methodology.
- Collection Development & Management - acquire, maintain, and manage a collection of data files used for secondary analysis by the local user community; purchase institutional data subscriptions; act as a site representative to data providers and national data archives for the institution.
- Preservation and Data Sharing Services - act on a strategy of preservation of datasets in the collection, such as media refreshment and file format migration; download and keep records on updated versions from a central archive. Also, assist users in preparing original data for secondary use by others; either for deposit in a central archive or institutional repository, or for less formal ways of sharing data. This may also involve marking up the data into an appropriate XML standard, such as the Data Documentation Initiative, or adding other metadata to facilitate online discovery

Source: http://en.wikipedia.org/wiki/Data_library [2013-10-21]

Oma käsitykseni on, että…

- Tutkimusdatan kerääjillä ei tule olemaan merkittävästi aikaa avatun datan jatkokäytön tukeen
- Sekä keskitetyissä datapalveluissa että tieteellisissä kirjastoissa työskentelevien datapalveluammattilaisten lukumäärä tulee kasvamaan merkittävästi 2010- ja 2020-luvuilla
- Alan koulutusta on järjestettävä valtakunnallisesti; asia on kiireellinen ja vaatii eritasoisia toimia
- Tutkimusaineistojen avoimuuden ja jatkokäytön tuki edellyttää tieteenala-asiantuntemusta
- Missä tahansa skenaariossa kirjastojen on lisättävä asiantuntemustaan jonkinasteisten datapalvelujen tuottajina

Keskustelua ja kysymyksiä